CIKM 2019 EComm AI:
Efficient User Interests Retrieval

CIKM2019 alimama 搜索推荐事业部

阿里巴巴算法大学 Alibaba Algorithm University

Alibaba Cloud | TIANCHI天池

QDU Team Profile

Chuanyu Xue
Pre-final Year Student at Qingdao University,
achieved top 3 rankings in three data mining competitions

Zhuoran Zhang
Algorithm Engineer in Spring Airlines,
achieved top 10 rankings in many data mining competitions

Shunyao Wu
Assistant Professor at Qingdao University,
achieved top 2 rankings in three data mining competitions

Problem Understanding

- **Task:**

  Predict the top-k preferred item from a large-scale item set for each user, under the liner complexity constraint.

- **Evaluation Metrics:**

$$Recall50(u) = \frac{|P_u \cap (G_u - H_u)|}{|G_u - H_u|}$$

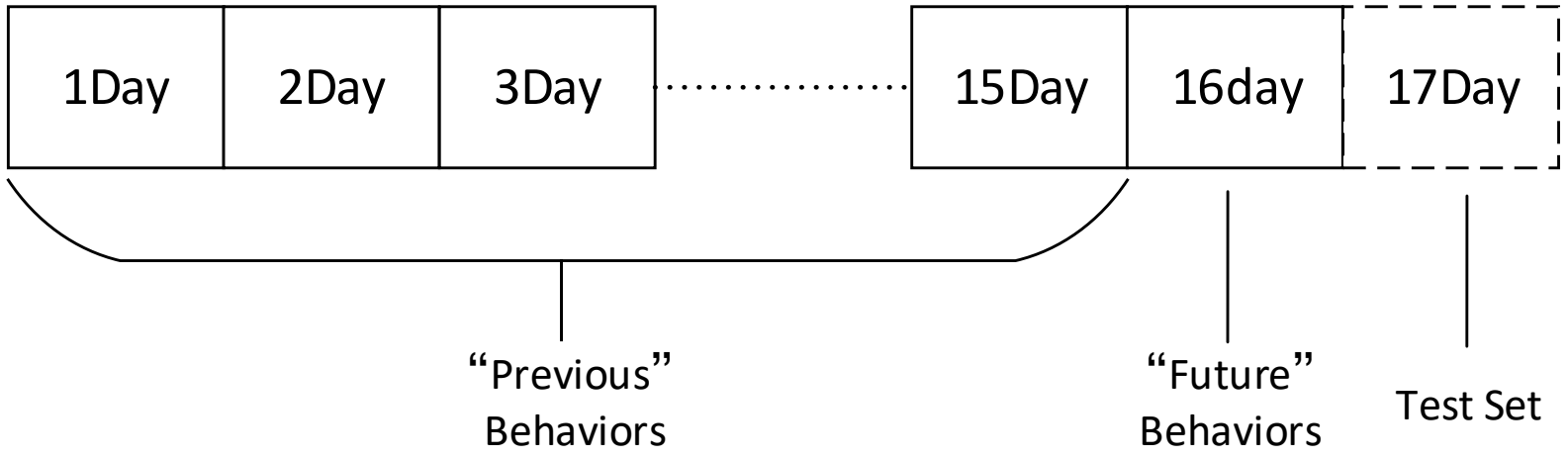
  P_u is recommendation set, G_u is ground truth, H_u is historical item set.

- **Given Dataset:**

  The training dataset has three files, including user behavior file, user profile file and item information file.

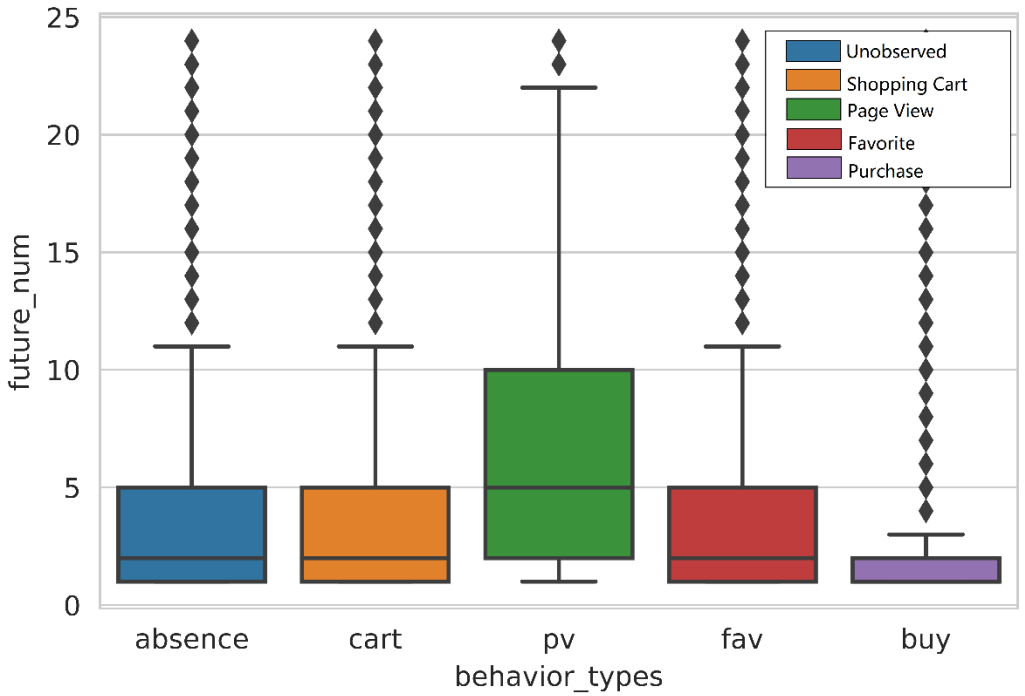
Exploratory Data Analysis

- **Some important points**
  1. Users' behavior types
  2. Time effect
  3. Category / Brand of items
  4. Popularity of items
- **Data Split**

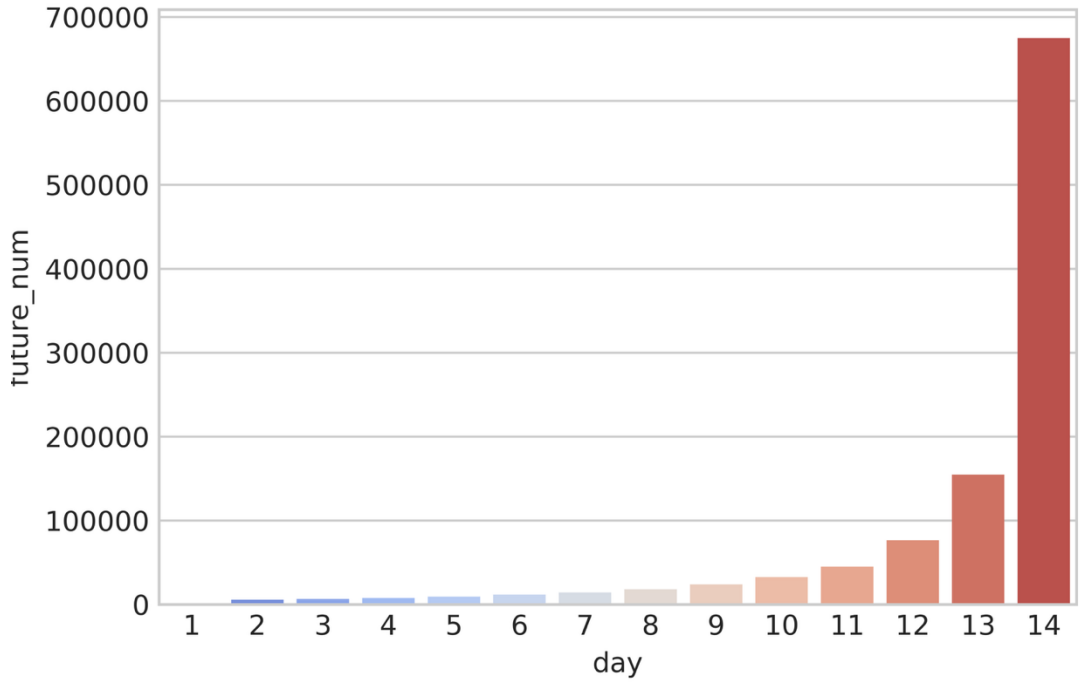


Exploratory Data Analysis

- **Implicit Feedback**



Previous PV will be more preferred,
Previous BUY will be less preferred.



Recent behaviors will have more influence

Exploratory Data Analysis

- **Implicit Feedback**

$$V_{u,i} = \max(s_{pv}x_{u,i},\ s_{fav}x_{u,i},\ s_{cart}\,x_{u,i},\ s_{buy}\,x_{u,i})$$

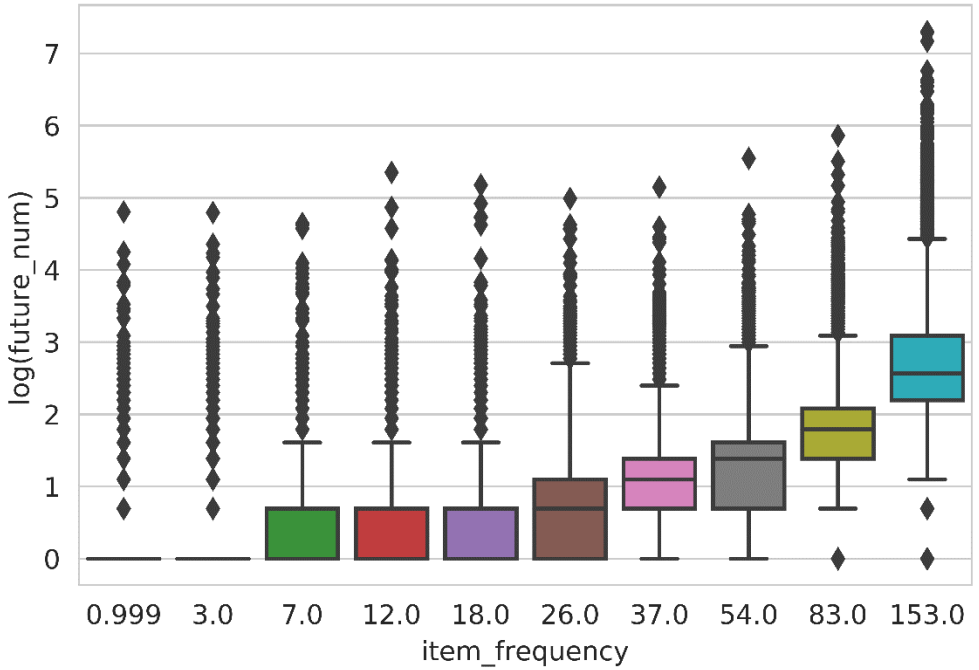
$$T_{u,i} = 1 - \left(\frac{D_{\max} - D_{u,i} + 1}{D_{max} - D_{min} + 1}\right)$$

$$R_{u,i} = T_{u,i} * V_{u,i}$$

$x_{u,i}$ is the behavior of user u to item i,
$s_{pv},\ s_{fav},\ s_{cart},\ s_{buy}$ are weights of different behaviors,
$D_{u,i}$: timestamp of behavior ($D_{u,i} = day + hour\ \%\ 24$)

Exploratory Data Analysis

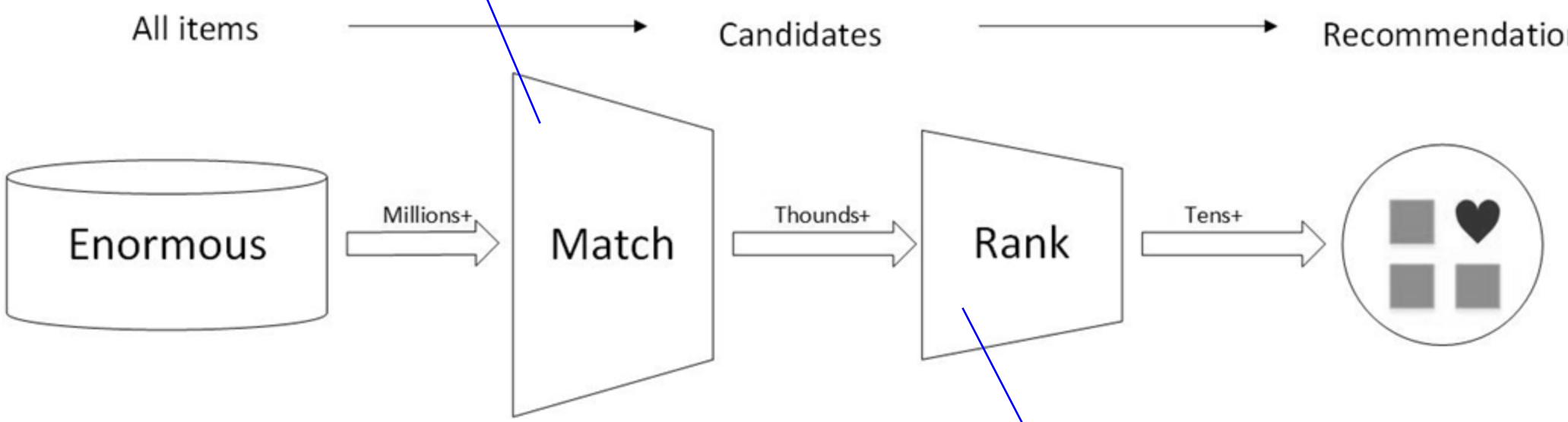
- **Popularity of items**



Popular items will be more preferred

Basic Idea of Recommendation

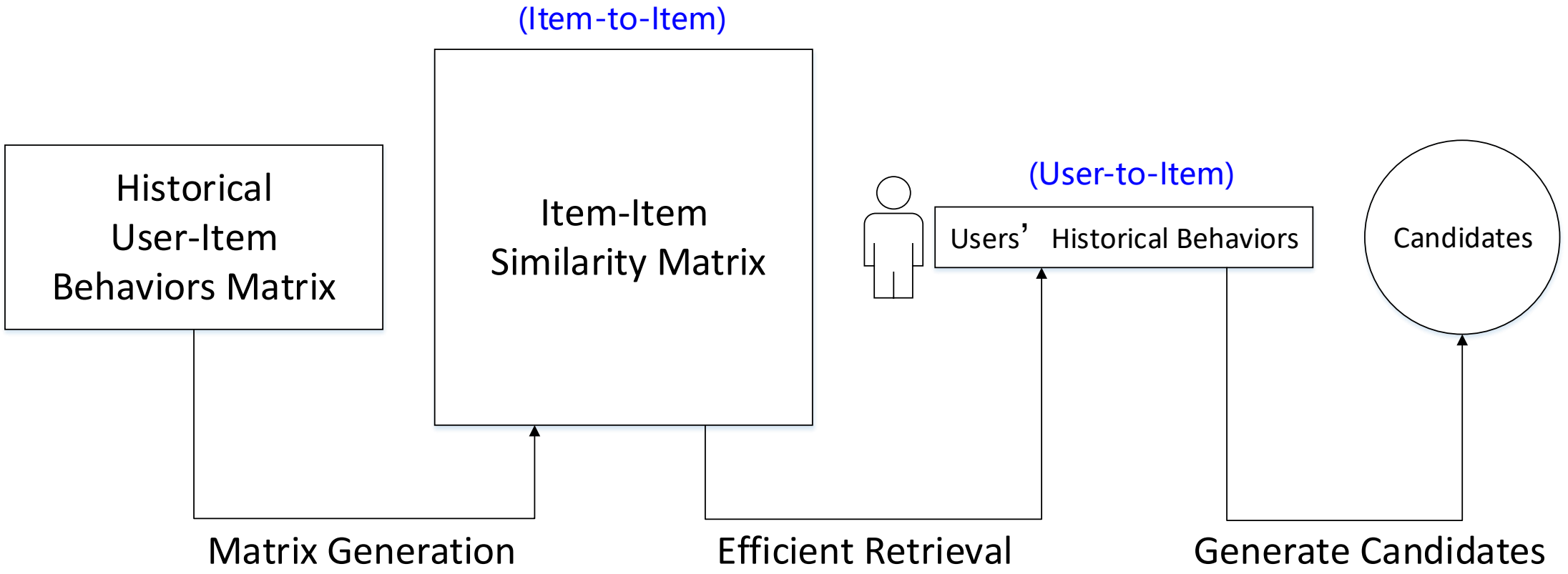
- **An Advanced Similarity for Item CF**
- **Parallel Algorithm and Data Structure for Efficient Matching**



- **A Distribution-Free Test of Independence for Feature Selection**
- **Liner and Power Model weighting method**

Matching

- **Main Process**



Matching

- **Association Rules**

$$\text{Confidence}(a,b) = P(b|a) = \frac{|U_a \cap U_b|}{|U_a|}$$

- **TF-IDF**

$$w_u = \frac{1}{\log(I_u) + 1}$$

Inactive users have more influence

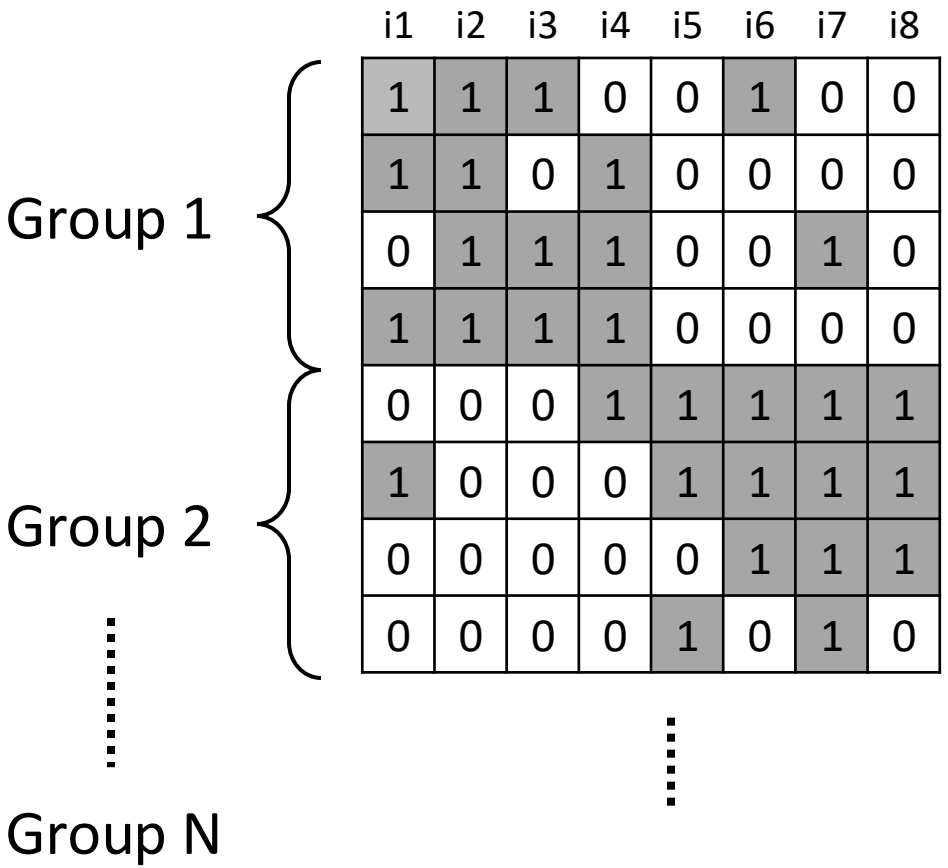
- **Advanced Similarity for Item-CF:**

$$Similarity(a,b) = \frac{\sum_{u \in U} w_u \delta(a,b)}{\sum_{u \in U_a} w_u}$$

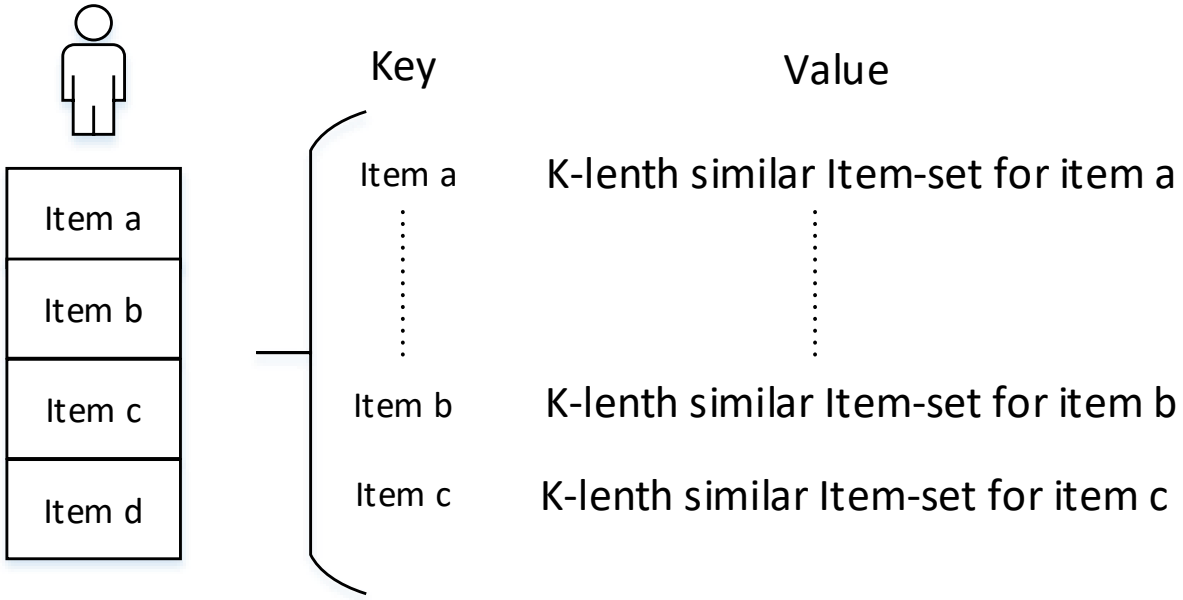
$$\delta(i,j) = \begin{cases} 1, & i \in I_u \; and \; j \in I_u \\ 0, & else \end{cases}, \quad when \; w_u \to 1, \text{Confidence(a, b)} = Similarity(\text{a}, b)$$

Matching

- **Parallel Algorithm for Similarity Matrix Generation**



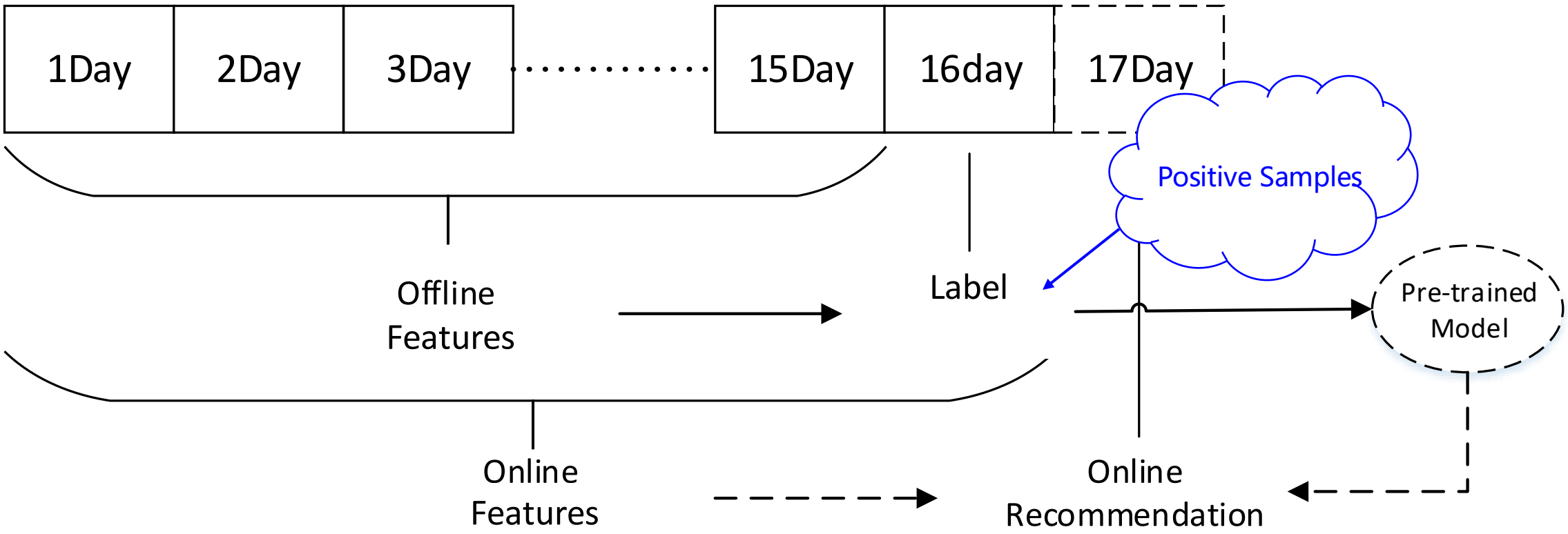
- **Data Structure for Efficient Retrieval**



Hash with only 430K values

Ranking

- **Main Process**



Ranking

- **Feature Engineering**
  - **Item Features**
    1. Statistic Values of item's ratings
    2. Timestamps of item's ratings
    3. Frequency of item
    4. Rank of item
  - **Category/Shop/Brand(item-set) Features**
    1. Statistic Values of item-set's ratings
    2. Timestamps of item-set's ratings
    3. Frequency of item-set
    4. Size of item-set
  - **User Interaction Features**
    1. Number of User's ratings on item-set
    2. Timestamps of user's ratings on item-set
    3. User's different behaviors on item-set
  - **Similarity Features**
    1. **Item's similarity**
    2. **Rank of item's similarity**

Ranking

- **Feature Selection**
  - **MV Test**: Mean Variance Test (***JASA** 2015)*
  - Distribution free test of Independence (https://github.com/ChuanyuXue/MVTest)
  - Mean Variance Index (X: a continuous r.v.; Y: a categorical one):

$$MV(X|Y) = E_X[Var_Y(F(X|Y))] \; where \; F(x|Y) = P(X \leq x|Y)$$

  - Testing hypothesis:

$$H_0: F_r(x) = F(x) \text{ for any } x \text{ and r=1,.. .,R}$$
$$H_1: F_r(x) \neq F(x) \text{ for some } x \text{ and r=1,.. .,R}$$
$$where, F(x) = P(X \leq x), F_r(x) = P(X \leq x|Y = y_r)$$

  - Test statistic:

$$T_n = n\widehat{MV}(X|Y)$$
$$= \sum_{r=1}^{R} \sum_{i=1}^{n} \widehat{p_r} * [\hat{F}_r(X_i) - \hat{F}(X_i)]^2$$

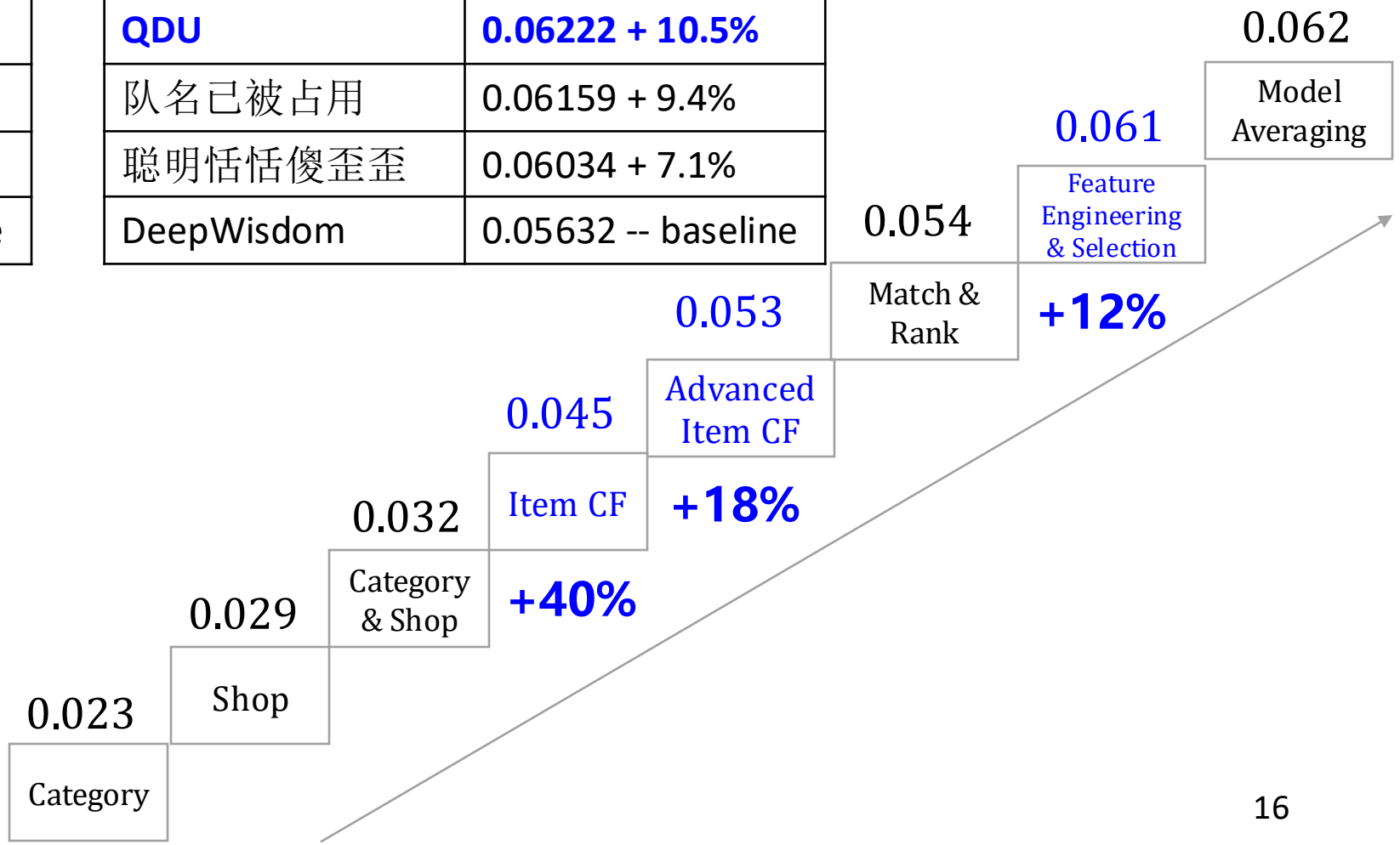
Ranking

- **Model Averaging**
  - **3 steps**
  - Step 1: averaging lightgbm and catboost with Harmonic Mean
  - Step 2: averaging lightgbm and catboost with Geometric Mean
  - Step 3: Harmonic Mean * 0.5 + Geometric Mean * 0.5

Conclusion

Team	The Qualification
QDU	**0.02645 + 6.5%**
聪明恬恬傻歪歪	0.02553 + 2.8%
去网吧里偷耳机	0.02543 + 2.4%
山有木兮	0.02516 + 1.2%
北方的郎	0.02484 -- baseline

Team	The Semi-Finals
江离	0.06246 + 10.9%
QDU	**0.06222 + 10.5%**
队名已被占用	0.06159 + 9.4%
聪明恬恬傻歪歪	0.06034 + 7.1%
DeepWisdom	0.05632 -- baseline



References

[1] Y. Huang et al. Tencentrec: Real-time stream recommendation in practice. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. 2015: 227-238.

[2] H. Cui et al. Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*. 2015, 110(510): 630-641.

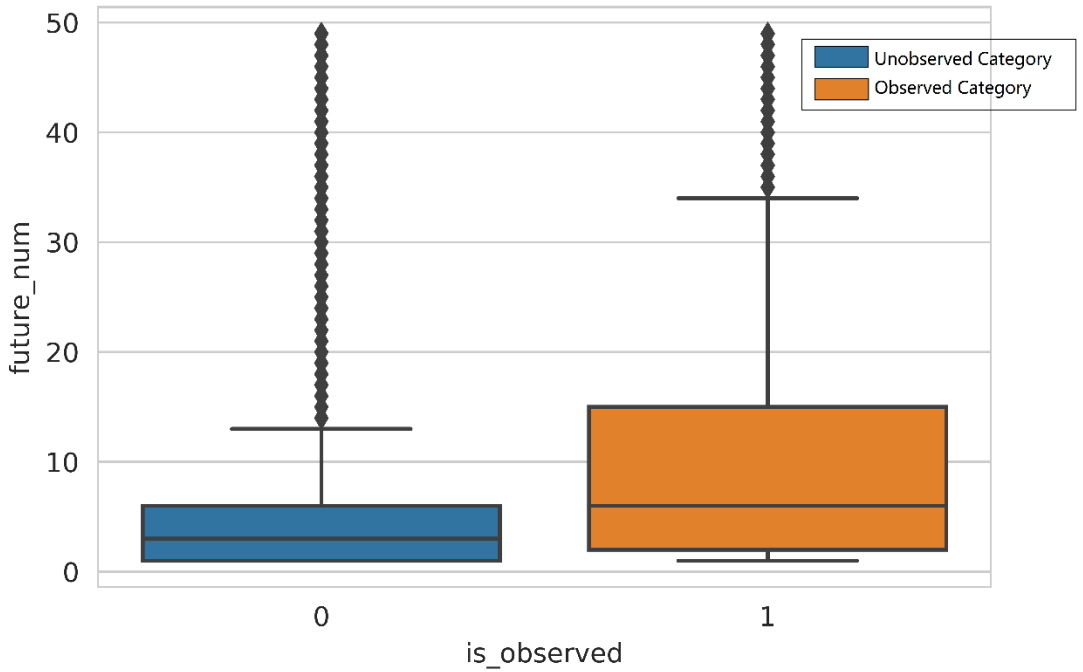
[3] H. Cui et al. A Distribution-Free Test of Independence and Its Application to Variable Selection. *arXiv preprint arXiv:1801.10559*, 2018.

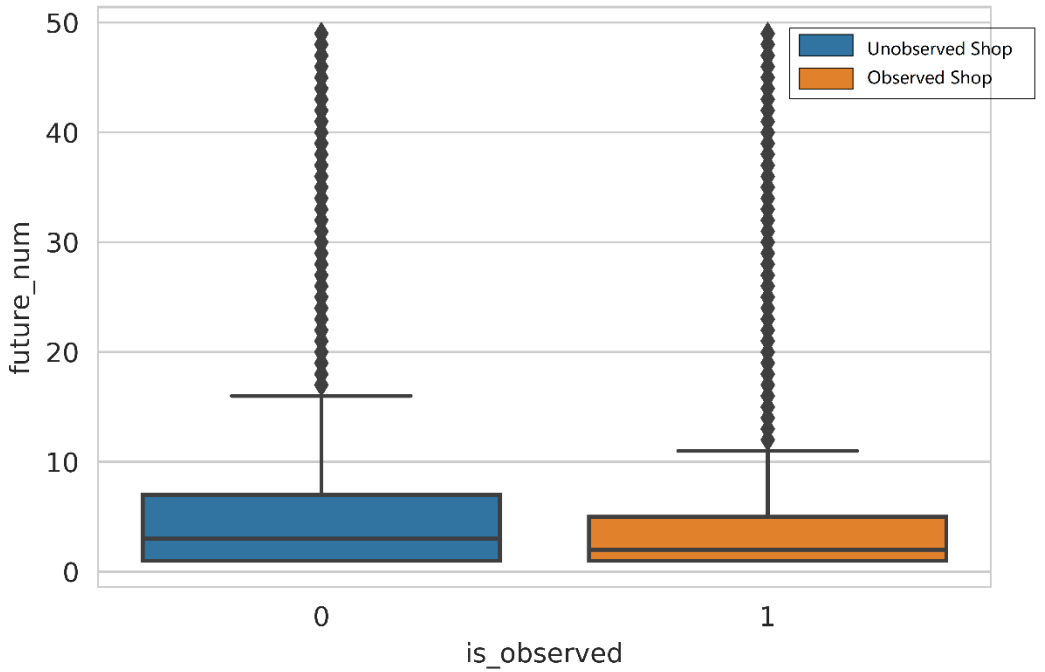
Thank You

← Contact me

Supplementary

- **Category / Shop of items**





Observed Categories will be more preferred

Observed Shops will be less preferred

Supplementary

- **User-to-Item**

➢ User historical behaviors:

Item	Apple	Banana	Football
Behavior	4.1	2.2	0.9

➢ Similarity Hash set:

Item	Item(Similarity)	Item(Similarity)	Item(Similarity)
Apple	Pineapple(0.9)	Pear(0.6)	Peach(0.4)
Banana	Mango(0.8)	Lemon(0.3)	
Football	Basketball(0.9)	Baseball(0.5)	

➢ Generated Candidates:

Top(500)(Pineapple(4.1 * 0.9) , Pear(4.1 * 0.6) , ……)

Supplementary

- **Ranking**

➢ User's Ground Truth in 16days

Item	Pineapple	Pear	Bicycle	Burger

➢ User's retrievaled items in 16days

Item	Pineapple	Pear	Mango	Lemon

➢ User's training samples

Item	Pineapple	Pear	Mango	Lemon	Bicycle	Burger
	Positive Sample	Positive Sample	Negative Sample	Negative Sample	Abandon	Abandon

Supplementary

- **Model Weighting**

$$result_{liner} = \frac{1}{(\frac{0.5}{result_{lgb}} + \frac{0.5}{result_{catb}})}$$

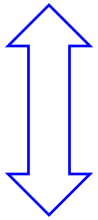
$$result_{power} = \sqrt{result_{lgb}{}^{0.5} * result_{catb}{}^{0.5}}$$

$$result = 0.5 * result_{liner} + 0.5 * result_{power}$$

where,
$result_{lgb}$ is result of LightGBM model,
$result_{catb}$ is result of Catboost model,

- **MV Test**
  - **Lemma 1**. *$MV(X|Y) = 0$ if and only if X and Y are statistically independent.*
  - *Test of Independence:*

$$H_0: X \text{ and } Y \text{ are statistically indep.}$$
$$H_1: X \text{ and } Y \text{ are not statistically indep.}$$



$$H_0: F_r(x) = F(x) \text{ for any } x \text{ and r=1,.. .,R}$$
$$H_1: F_r(x) \neq F(x) \text{ for some } x \text{ and r=1,.. .,R}$$